Zur Analyse von Paneldaten mit SPSS/PC: die EGLS-Schätzung des Fehlerkomponentenmodells

Kühnel, Steffen M.

Zur Analyse von Paneldaten mit SPSS/PC: Die EGLS-Schätzung des Fehlerkomponentenmodells

Ein Nachtrag zum Frühjahrsseminar 1992

von Steffen M. Kühnel

Bei Paneldaten werden interessierende Merkmale von Untersuchungseinheiten im Zeitverlauf mehrfach erhoben. Bei der Analyse solcher Daten mit Regressionsmodellen stellt sich das Problem, daß als Folge der Meßwiederholungen mit korrelierten Residuen zu rechnen ist. Statistische Modelle zur Panelanalyse berücksichtigen dies. Ein Modell, das u.a. im Frühjahrsseminar 1992 zur Analyse zeitbezogener Daten vorgestellt wurde, ist das Fehlerkomponentenmodell. ***U. Rendtel*** (1992) hat in seiner Vorlesung zur Panelanalyse gezeigt, wie dieses Modell in den Programmsystemen SAS oder GAUSS über mehrfache einfache Kleinstquadratschätzungen berechnet werden kann. Ich möchte in diesem Beitrag auf die Berechnung des Modells mit SPSS/PC eingehen. Zuvor werde ich kurz auf das statistische Modell und die Idee der verwendeten Schätzmethode eingehen. Dabei werde ich auf formale Darstellungen in Matrizenschreibweise sowie auf mathematisch-statistische Ableitungen verzichten. Diese sind etwa bei ***Hübler*** (1990) und bei ***Rendtel*** (1992) zu finden.

1. Das Fehlerkomponentenmodell zur Analyse von Paneldaten

In der quantitativen Sozialforschung erfolgt die Analyse des Zusammenhangs zwischen einer abhängigen Variable Y und erklärenden Variablen $X_1, X_2, ..., X_k$ oft über ein lineares Regressionsmodell. Das Modell postuliert, daß für alle Fälle i=1, 2, ..., n die Beziehung zwischen den Werten y_i der abhängigen Variable und den Werten x_{ji} der Prädiktoren als lineare Gleichung dargestellt werden kann:

$$(1) \qquad \begin{aligned} y_i &= \alpha + x_{1i}\cdot\beta_1 + x_{2i}\cdot\beta_2 + \ldots + x_{ki}\cdot\beta_k + u_i \\ &= \alpha + \sum_{j=1}^{k} x_{ji}\cdot\beta_j + u_i \end{aligned}$$

Das statistische Modell geht in der Regel davon aus, daß die Residuen u_i in Gleichung (1) voneinander unabhängige identisch verteilte Zufallsvariablen sind, deren Erwartungswerte $\mu(u_i)$ stets Null sind. Ziel der statistischen Analyse ist es dann u.a., die Regressionskoeffizienten $\alpha, \beta_1, \beta_2, ..., \beta_k$ möglichst optimal zu schätzen.

Bei Panelanalysen liegen für jeden Fall mehrere zeitversetzte Beobachtungen vor. Formal läßt sich dies so darstellen, daß die Variablen in Gleichung (1) um Zeitindizes t ergänzt werden, die den Meßzeitpunkt t der Beobachtung identifizieren. Die Gleichung hat dann folgende Form:

(2) $$y_{it} = \alpha + \sum_{j=1}^{k} x_{jit} \cdot \beta_j + u_{it}$$

Bei der Berücksichtigung des Meßzeitpunktes ist es sinnvoll, zwischen zeitveränderlichen und zeitkonstanten erklärenden Variablen zu unterscheiden. Zeitveränderliche Variablen können zu jedem Meßzeitpunkt einen anderen Wert aufweisen, zeitkonstante Variablen variieren dagegen nicht über die Zeit. Typische Beispiele für zeitveränderliche abhängige Variablen sind in der Umfrageforschung die Antworten auf Einstellungsfragen. Das Geschlecht eines Befragten ist dagegen eine zeitkonstante Variable: unabhängig vom Meßzeitpunkt ist ein Befragter immer weiblich oder männlich. Um zwischen zeitveränderlichen und zeitkonstanten erklärenden Variablen unterscheiden zu können, wird im folgenden die Gesamtheit der erklärenden Variablen in die Teilmenge $\{X_1, X_2, ..., X_p\}$ der p zeitveränderlichen Prädiktoren und in die Teilmenge $\{Z_1, Z_2, ..., Z_k\}$ der zeitkonstanten Prädiktoren zerlegt. Die Regressionskoeffizienten der zeitveränderlichen Variablen werden durch das Zeichen "β" symbolisiert, die Regressionskoeffizieten der zeitkonstanten Variablen dagegen durch das Symbol "γ". Da bei zeitkonstanten Variablen der Zeitindex t ausgelassen werden kann, wird Gleichung (2) zu:

(3) $$y_{it} = \alpha + \sum_{j=1}^{p} x_{jit} \cdot \beta_j + \sum_{l=1}^{k} z_{li} \cdot \gamma_l + u_{it}$$

Versucht man nun, mit Hilfe der einfachen Kleinstquadratmethode die Regressionskoeffizienten aus Gleichung (3) zu schätzen, so ist mit ineffizienten Ergebnissen zu rechnen, wenn die Residuen u_{it} einer Person i nicht unabhängig voneinander sind. So ist es etwa denkbar, daß ein Befragter relativ gesehen zu den übrigen Befragten, unabhängig von den Werten der Prädiktoren, die Tendenz haben kann, bei der abhängigen Variable immer etwas höhere Werte anzugeben. Wegen der daraus folgenden Verletzung der Homoskedastizitätsannahme werden bei einer einfachen Kleinstquadratschätzung (OLS-Regression) der Regressionskoeffizienten die Standardfehler verzerrt geschätzt. Infolgedessen sind auch die inferenzstatistischen Tests ungültig.

Um diese Fehlerquelle zu beseitigen, ist es nötig, die Autokorrelationen zwischen den Meßwiederholungen zu berücksichtigen. Dies geschieht dadurch, daß der nichterklärte Rest u_{it} der Regression der abhängigen Variable in zwei Komponenten zerlegt wird:

(4) $$u_{it} = \delta_i + \varepsilon_{it}$$

In der Gleichung steht δ_i für eine Residualkomponente, die über die Zeit konstant ist, also z.B. für die Tendenz, stets einen etwas höheren Wert bei der abhängigen Variable anzugeben. Bei der zweiten Komponente ε_{it} wird dagegen angenommen, daß sie zu jedem Meßzeitpunkt einen unterschiedlichen Wert aufweisen kann. Aufgrund der Aufteilung der Residuen in zwei Komponenten wird dieses Modell in der Literatur als **Fehlerkomponentenmodell** bezeichnet.[1] Zur Abkürzung verwende ich im folgenden die Bezeichnung **EC-Modell,** wobei 'EC' für 'error component' steht.

Da es sich bei beiden Komponenten um Residuen handelt, wird unterstellt, daß jede Komponente einen Erwartungswert von Null aufweist und daß jede Komponente auch von den erklärenden Variablen des Modells unabhängig ist. Weiter wird davon ausgegangen, daß die einzelnen Fälle unabhängig sind. Aufgrund der Unabhängigkeit der zeitvariablen Komponente ε_{it} von δ_i gilt dann für die Grundgesamtheitsmittelwerte (Erwartungswerte) $\mu(u_{it})$, Varianzen $\sigma^2(u_{it})$ und Kovarianzen $\sigma(u_{it},u_{i't'})$ der ursprünglichen Residuen u_{it}:

(5)
$$\begin{aligned}
\mu(u_{it}) &= \mu(\delta_i) = \mu(\varepsilon_{it}) = 0 \\
\sigma^2(u_{it}) &= \sigma_\delta^2 + \sigma_\varepsilon^2 \\
\sigma(u_{it},u_{it'}) &= \sigma_\delta^2 \qquad t \neq t' \\
\sigma(u_{it},u_{i't'}) &= 0 \qquad i \neq i'
\end{aligned}$$

Wenn die Varianzen σ_δ^2 und σ_ε^2 der Fehlerkomponenten bekannt wären, ließen sich die Regressionskoeffizienten mit Hilfe des verallgemeinerten Kleinstquadratschätzers (GLS-Schätzer) unverzerrt und effizient schätzen. Tatsächlich kann die Schätzung in zwei Schritten erfolgen. Im ersten Schritt werden die Varianzen der beiden Fehlerkomponenten geschätzt. In einem zweiten Schritt werden diese Schätzwerte dann in der verallgemeinerten Kleinstquadratschätzung der Regressionskoeffizienten eingesetzt. Diese Schätzmethode wird als **EGLS-Schätzung** bezeichnet, wobei der Ausdruck "EGLS" für "estimated gene-

[1] Es gibt auch komplexere Modelle, in denen zusätzlich für die Meßzeitpunkte eine Fehlerkomponente modelliert wird (vgl. *Hübler*, 1990: 70).

ralized least squares" steht. Der Vorteil dieser Schätzmethode besteht darin, daß sie durch mehrfache Anwendung gewöhnlicher Regressionsrechnungen bewerkstelligt werden kann.

2. EGLS-Schätzung über mehrfache einfache Regressionen

Im ersten Schritt der EGLS-Schätzung müssen die Varianzen der Fehlerkomponenten geschätzt werden. Dies erfolgt über zwei einfache Regressionen. Zunächst wird für jeden Fall i der Stichprobe der Mittelwert der Variablen über die Zeit gebildet. Aus dem EC-Modell folgt dann für die Mittelwerte:

(6)
$$\bar{y}_i = \frac{1}{T_i}\sum_{t=1}^{T_i} y_{it} = \alpha + \sum_{j=1}^{p} \bar{x}_{ji}\cdot\beta_j + \sum_{l=1}^{K} z_{li}\cdot\gamma_l + \bar{u}_i$$
$$\text{mit:} \quad \bar{u}_i = \delta_i + \bar{\varepsilon}_i$$

In Gleichung (6) steht T_i für die Anzahl der Beobachtungen eines Falles i über die Zeit. Es ist nicht notwendig, daß alle Fälle gleich oft beobachtet werden. Zieht man die Mittelwerte über die Zeit von den ursprünglichen Werten ab, kürzen sich die zeitkonstanten Variablen und Koeffizienten heraus. Das Modell reduziert sich zum sogenannten **Within-Modell:**

(7)
$$(y_{it} - \bar{y}_i) = \sum_{j=1}^{p} (x_{jit} - \bar{x}_{ji})\cdot\beta_j + (\varepsilon_{it} - \bar{\varepsilon}_i)$$
$$\tilde{y}_i = \sum_{j=1}^{p} \tilde{x}_{ji}\cdot\beta_j + \tilde{\varepsilon}_i$$

Zu beachten ist, daß Gleichung (7) keine Regressionskonstante enthält. Es läßt sich nun zeigen, daß die Regression von $\tilde{y}$ auf die Prädiktoren $\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_p$ nach der gewöhnlichen Kleinstquadratmethode (ohne Interzept) zu erwartungstreuen und konsistenten Schätzern $\beta_{(w)1}, \beta_{(w)2}, \ldots, \beta_{(w)p}$ der Regressionskoeffizienten $\beta_1, \beta_2, \ldots, \beta_p$ führt. Darüber hinaus ergibt sich als erwartungstreuer Schätzer für die Varianz σ_ε^2:[2]

[2] Der Abzug der zeitbezogenen Mittelwerte von den Ausgangsdaten läßt sich in Matrizenschreibweise als eine Multiplikation von links mit einer blockdiagonalen Matrix darstellen, wobei jeder Block eine idempotente-symmetrische Submatrix ist. Obwohl die Residuen der Regression der transformierten Daten nicht homoskedastisch sind, ist der Erwartungswert der Summe der quadrierten Residuen der Kleinstquadratregression hier proportional zur Varianz σ_ε^2.

(8) $$\hat{\sigma}_\varepsilon^2 = \frac{\sum_{i=1}^{n} \left(\bar{y}_i - \sum_j \bar{x}_{ji} \cdot \beta_{(W)j} \right)^2}{\sum_{i=1}^{n} T_i - n - p}$$

Über das Within-Modell kann also die Varianz einer der beiden Fehlerkomponenten geschätzt werden. Die Vermutung liegt nahe, daß die Varianz der anderen Fehlerkomponente über die Regression der Mittelwerte aus Gleichung (6) ermittelt werden kann. Dieses Regressionsmodell über die Mittelwerte wird als **Between-Modell** bezeichnet. Betrachten wir dazu die Residuen dieser Gleichung, so gilt:

(9) $$\sigma^2(\bar{u}_i) = \sigma_\delta^2 + \frac{1}{T_i} \cdot \sigma_\varepsilon^2$$

Bei über die Fälle variierenden T_i sind die Residuen heteroskedastisch: Je nach der Anzahl der Beobachtungen eines Falles über die Zeit ist die Varianz $\sigma^2(\bar{u}_i)$ unterschiedlich groß. Faßt man dagegen Fälle mit gleicher Anzahl von Beobachtungen zusammen, so haben in dieser Teilgruppe die Residuen der Regression über die Mittelwerte die gleiche Varianz. Schätzt man für eine solche Teilgruppe, bei denen jeweils T Beobachtungen vorliegen, die Between-Regressionskoeffizienten $\hat{\alpha}_{(B)}$, $\hat{\beta}_{(B)1}$, $\hat{\beta}_{(B)2}$, ..., $\hat{\beta}_{(B)p}$, $\hat{\gamma}_{(B)1}$, $\hat{\gamma}_{(B)2}$, ..., $\hat{\gamma}_{(B)k}$ aus Gleichung (6) mit der üblichen Kleinstquadratmethode, gilt[3] für die beobachteten Residuen dieser Regression:

(10) $$\frac{\mu\left(\sum_{i=1}^{n} \left(\bar{y}_i - \hat{\alpha}_{(B)} - \sum_j \bar{x}_{ji} \cdot \hat{\beta}_{(B)j} - \sum_l z_{li} \cdot \hat{\gamma}_{(B)l} \right)^2 \right)}{n - p - k - 1} = \frac{\mu\left(\sum_{i=1}^{n} \bar{e}^2 \right)}{n - p - k - 1} = \sigma_\delta^2 + \frac{1}{T}\sigma_\varepsilon^2$$

Ersetzt man in Gleichung (10) den Erwartungswert der Summe der beobachteten quadrierten Residuen durch die Summe selber und setzt für σ_ε^2 den Schätzwert aus Gleichung (8) ein, ergibt sich ein konsistenter Schätzer für die zweite Fehlervarianzkomponente:

---

[3] In Gleichung (10) und (11) ist zu beachten, daß sich die Fallzahl n hier nur auf die Fälle bezieht, für die genau T Beobachtungen vorliegen.

(11)
$$\hat{\sigma}_\delta^2 = \frac{\sum_{i=1}^{n} \bar{e}^2}{n-p-k-1} - \frac{1}{T}\hat{\sigma}_\varepsilon^2$$

Damit ist der erste Schritt der EGLS-Schätzung abgeschlossen. Im zweiten Schritt werden die geschätzten Fehlervarianzkomponenten für die verallgemeinerte Kleinstquadratschätzung genutzt. Anstelle einer direkten verallgemeinerten Kleinstquadrat-Schätzung werden die Ursprungsdaten so transformiert, daß die Residuen homoskedastisch werden. Die übliche Kleinstquadratschätzung über die so transformierten Daten ergibt die gleichen Ergebnisse wie eine direkte verallgemeinerte Kleinstquadratschätzung.

Für die Transformation sind zunächst die Werte ϑ_i einer Skalierungsvariable zu berechnen:

(12)
$$\vartheta_i = 1 - \sqrt{\frac{\hat{\sigma}_\varepsilon^2}{\hat{\sigma}_\varepsilon^2 + T_i \cdot \hat{\sigma}_\delta^2}}$$

Anschließend wird in den Ausgangsdaten von allen Variablen, einschließlich der Konstante "Eins" zur Schätzung des Interzepts, der jeweilige mit ϑ_i multiplizierte Mittelwert über die Zeit abgezogen:

(13)
$$\begin{aligned} q_{it} &= y_{it} - \vartheta_i \cdot \bar{y}_i \\ c_i &= 1 - \vartheta_i \\ v_{jit} &= x_{jit} - \vartheta_i \cdot \bar{x}_{ji} \qquad j = 1,2,\ldots,p \\ w_{li} &= z_{li} - \vartheta_i \cdot z_{li} \qquad l = 1,2,\ldots,k \end{aligned}$$

Die Berechnung der gewöhnlichen Kleinstquadratlösung (ohne Interzept) der Regressionsgleichung:

(14)
$$q_{it} = c_i \cdot \hat{\alpha} + \sum_{j=1}^{p} v_{jit} \cdot \hat{\beta}_j + \sum_{l=1}^{p} w_{li} \cdot \hat{\gamma}_l + e_{it}$$

ergibt schließlich die gesuchten EGLS-Schätzer der ursprünglichen Regressionsgleichung (3). Asymptotisch korrekt sind auch die dabei mitberechneten Standardfehler und T-Werte der Koeffizienten.

3. Die Berechnung der EGLS-Lösung mit SPSS/PC

Die Darstellung der EGLS-Schätzmethode im vorherigen Abschnitt zeigt, daß alle Berechnungen mit einfachen Kleinstquadratschätzungen multipler Regressionsgleichungen durchgeführt werden können. Das EC-Modell kann somit mit beliebigen Statistikprogrammen geschätzt werden. Im folgenden möchte ich darstellen, wie die Berechnung in SPSS/PC erfolgt.

Betrachtet man die Ausgangsgleichung (3) des EC-Modells, so fällt auf, daß Messungen einer Größe zu verschiedenen Zeitpunkten als *eine* doppelt indizierte Variable aufgefaßt werden. Oft sind die Paneldaten in einer Datei allerdings so angeordnet, daß die mehrfachen Messungen einer Größe zu verschiedenen Zeitpunkten als verschiedene Variablen abgespeichert sind. Wenn dies der Fall ist, müssen die Variablen zunächst rearrangiert werden. Dazu werden die relevanten Variablen jeder Panelwelle in temporäre Systemdateien gespeichert und anschließend mit dem SPSS/PC-Kommando "JOIN ADD" so zusammengefügt, daß die Meßwiederholungen unter einem Variablennamen gespeichert werden. Zeitkonstante Variablen werden dabei mehrfach herausgeschrieben. Zur späteren Identifikation werden außerdem die Fallnummern und eine Variable, die den Meßzeitpunkt enthält, in die temporären Dateien herausgeschrieben. Außerdem empfiehlt es sich, auf diese Weise gleich ungültige Fälle auszuschließen.

Wenn "idnr" der Variablenname zur Identifikation der Fallnummer, "yl" die abhängige Variable zum ersten Meßzeitpunkt sei, "xll" und "x21" zwei zeitveränderliche erklärende Variablen zu diesem Meßzeitpunkt und "z" eine zeitkostante erklärende Variable, würden die SPSS-Anweisungen zum Herausschreiben der Daten des ersten Meßzeitpunkts folgendermaßen aussehen:

```
COUNT nmissing=yl xll x21 z (MISSING).
COMPUTE zeit=l.
PROCESS IF (nmissing EQ 0).
SAVE OUT="templ.sys"
    /KEEP=idnr zeit yl xll x21 z /RENAME (yl xll x21 z=y xl x2 z).
```

Analog würden auch die Daten der übrigen Meßzeitpunkte in temporäre Dateien herausgeschrieben. Wichtig ist, daß die herausgeschriebenen Daten, die sich auf eine Größe beziehen, den selben Variablennamen erhalten. Dies wird durch die Spezifikation "RENAME .." im SAVE-Kommando erreicht. Die temporären Dateien werden dann mit dem Kommando "JOIN ADD" so zu einer Datei zusammengesetzt, daß alle Meßwerte einer Größe unabhängig vom Meßzeitpunkt in einer Variable stehen:

```
JOIN ADD /FILE="templ.sys" /FILE="ternp2.sys" /FILE="temp3.sys".
SORT CASES BY idnr zeit.
```

Im Beispiel werden die Daten eines dreiwelligen Panels zusammengefügt. Das Kommando "SORT CASES" sortiert anschließend die Daten nach Fallnummer und Meßzeitpunkt.

Für die Berechnung der Between- und Within-Regressionen müssen die Mittelwerte über die Zeit gebildet und in einer Datei abgespeichert werden. Die Mittelwertsbildung erfolgt mit dem Kommando "AGGREGATE":

```
AGGREGATE OUT=*
          /PRES  /BREAK=idnr
          /yq=MEAN(y)
          /xlq=MEAN(xl)
          /x2q=MEAN(x2)
          /z=MEAN(z)
          /ti=NU(y).
SAVE OUT="tempm.sys"  /COMP.
```

Das Kommando "AGGREGATE" faßt über den Unterbefehl "BREAK=idnr" die Werte eines Falles zusammen. Die Berechnung der Mittelwerte der abhängigen Variable über die Meßzeitpunkte wird durch den Unterbefehl "yq=MEAN(y)" erreicht. Analog werden die Mittelwerte der anderen Variablen berechnet. Für die Within-Regression sind die Ausgangsdaten um diese Mittelwerte zu bereinigen. Dazu müssen die Mittelwerte zunächst fallweise mit den Ausgangsdaten verknüpft werden. Dies geschieht mit dem "JOIN MATCH"-Kommando:

```
JOIN MATCH  TABLE="tempm.sys"
            /FILE="templ.sys"
            /BY idnr.
COMPUTE yw=y-yq.
COMPUTE xlw=xl-xlq.
COMPUTE x2w=x2-x2q.
```

Die Mittelwerte (Aggregatdaten) sind in der Datei "tempm.sys" gespeichert, die Ausgangsdaten (Individualdaten) in der Datei "templ.sys". Durch die Spezifikation "TABLE=..." wird dem SPSS-Programm mitgeteilt, daß eine hierarchische Verknüpfung der Daten erfolgt, so daß ein Fall in der Aggregatdatendatei mehreren Fällen in der Individualdatendatei zugeordnet werden kann. Die Verknüpfung erfolgt über die Fallnummer, was dem Programm durch den Unterbefehl "BY idnr" mitgeteilt wird. Nachdem die Mittelwerte mit den Ursprungsdaten verknüpft sind, werden in den nachfolgenden COMPUTE-Anweisungen die Mittelwerte von den Ausgangsdaten abgezogen. Die OLS-Regression über diese Daten ergibt die WITHIN-Schätzung:

```
REGRESSION VAR yw xlw x2w
           /ORIGIN /DEP=yw /ENT
           /SAVE=RESID(wresid).
COMPUTE wres2=wresid*wresid.
```

Bei der Regression ist zu beachten, daß keine Regressionskonstante geschätzt wird. In SPSS wird dies über die Spezifikation "ORIGIN" sichergestellt. Mit dem Unterbefehl "SAVE=RESID(wresid)" werden die Residuen der Within-Regression als zusätzliche Variable gespeichert. Die nachfolgende COMPUTE-Anweisung quadriert diese Residuen. Die Summe der Variable "wres2" über alle Beobachtungen ist der Zähler auf der rechten Seite von Gleichung (8). Die Division durch die Gesamtzahl der Beobachtungen minus der Fallzahl und der Anzahl der zeitvariablen Prädiktoren ergibt die Varianz σ_ε^2.

Die Between-Regression erfolgt über die aggregierten Mittelwerte. Durch das Kommando "SELECT IF (ti=t)" vor der Regression wird sichergestellt, daß nur Fälle mit gleicher Beobachtungszahl t in die Rechnung eingehen:[4]

```
SELECT IF (ti = t) .
REGRESSION VAR yq xlq x2q zq
           /DEP yq
           /ENT
           /SAVE=RESID(bresid).
COMPUTE bres2=bresid*bresid.
```

Wie bei der Within-Regression werden auch hier wieder die Residuen gespeichert und in der nachfolgenden COMPUTE-Anweisung quadriert. Teilt man die Summe der Werte dieser Variable durch die in der Regression eingehenden Fälle mit Beobachtungszahl t minus der Anzahl der in der Between-Regression geschätzten Regressionskoeffizienten und zieht davon die durch t geteilte Varianz σ_ε^2 ab, ergibt sich die Schätzung für die zweite Varianzkomponente σ_δ^2. Der Wert dieser Varianzen sei in den Variablen "sigma2e" und "sigma2d" gespeichert. Mit einer COMPUTE-Anweisung läßt sich dann die Skalierungsvariable ϑ nach Gleichung (12) berechnen. Anschließend erfolgt die Transformation der Ausgangsdaten entsprechend Gleichung (13) und schließlich die Regression zwischen den transformierten Variablen:

```
COMPUTE theta=l-SQRT(sigma2e/(sigma2e+sigma2d*ti)).
COMPUTE constt=l-theta.
COMPUTE yt=y-theta*yq.
COMPUTE xlt=xl-theta*xlq.
COMPUTE x2t=x2-theta*x2q.
COMPUTE zt=z-theta*z.
REGRESSION VAR=yt constt xlt x2t zt
           /ORIGIN /DEP=yt /ENT.
```

Das Ergebnis dieser letzten Regression ist die gesuchte EGLS-Lösung.

---

[4] Die Größe "t" muß natürlich vorher festgelegt und mit einer COMPUTE-Anweisung generiert worden sein.

4. Ein Anwendungsbeispiel

Im Anwendungsbeispiel soll die Haltung zur Volkszählung 1987 durch die Beurteilung der Notwendigkeit der Zählung, das politische Interesse und das Geschlecht des Befragten erklärt werden. Die Daten stammen aus dem dreiwelligen Panel der Kölner Begleituntersuchung zur Volkszählung.[5] Die Haltung zur Volkszählung wurde auf einer siebenstufigen Antwortskala erhoben, das politische Interesse auf einer fünfstufigen Skala. Die Notwendigkeit der Volkszählung ist dichotomisiert. Die Variablennamen sind der Notation dieses Beitrages entsprechend angepaßt. So bezeichnet "Yl" die Haltung zur Volkszählung in der ersten Welle, "Y2" die Haltung in der zweiten Welle und "Y3" die Haltung in der dritten Welle. "XU" enthält die Antworten für das politische Interesse zum ersten Meßzeitpunkt, "X12" und "X13" die Messungen in den Folgewellen. "X21", "X22" und "X23" stehen für die Beurteilung der Notwendigkeit der Zählung zum ersten, zweiten bzw. dritten Meßzeitpunkt. Das Geschlecht wird als zeitkonstante Variable "Z" genannt. Neben den Werten dieser Variablen enthält die SPSS-Systemdatei "ECBEISP.SYS" noch die Fallnummer (Variablenname: "IDNR").

Die nachfolgenden Tabellen zeigen, daß aus der ersten Welle 2514 Beobachtungen der zu analysierenden Variablen vorliegen, aus der zweiten Welle 1380 Beobachtungen und aus der dritten Welle 1254 Beobachtungen. Der Tabelle mit den Häufigkeit der Meßwiederholungen ist zu entnehmen, daß 1056 Personen nur einmalig befragt wurden, 312 Personen zweimal befragt wurden und 1167 Personen in allen drei Wellen befragt wurden.[6]

Beobachtungen pro Erhebungszeitpunkt	
1. Welle	2514
2. Welle	1380
3. Welle	1254
insgesamt:	5148

Anzahl Meßwiederholungen	
keine	1056
eine	312
zwei	1167

Die Berechnung der EGLS-Lösung erfolgt mit dem im Anhang dokumentierten SPSS-Befehlen. Im ersten Schritt werden die Daten so rearrangiert, daß die Messungen einer

[5] Der Datensatz ist unter der Studiennummer ZA1592 im Zentralarchiv archiviert. Zur Beschreibung der Studie vgl. *Scheuch u.a.,* 1989.

[6] Von den 312 Personen, die einmal wiederbefragt wurden, sind 225 Personen in der 1. und 2. Welle befragt worden und 87 Personen in der 1. und 3. Welle.

Größe über die Meßzeitpunkte hinweg in einer Variable stehen. Anschließend werden in den Variablen "P" und "K" die Anzahl der zeitveränderlichen und zeitkonstanten Prädiktoren festgehalten. Der Wert 3 in der Variable "TMAX" legt fest, daß später bei der Between-Regression nur Fälle berücksichtigt werden, bei denen drei Beobachtungen vorliegen. Ab dieser Stelle erfolgt die gesamte Berechnung automatisch. Die im Laufe der Berechnung notwenigen Hilfsgrößen zur Berechnung der Varianzkomponenten werden während der Schätzung ermittelt und zwischengespeichert.

```
Page  27   EGLS-Schätzung des 'Error-Component-Models' (EC-Modell)      5/7/92

Berechnung 'between'-Regression' über vollständige Fälle

        * * * *  M U L T I P L E   R E G R E S S I O N  * * * *

Equation Number 1    Dependent Variable..   YQ

Multiple R            .69961
R Square              .48945
Adjusted R Square     .48814
Standard Error       1.13089

Analysis of Variance
                 DF        Sum of Squares        Mean Square
Regression        3           1425.92970          475.30990
Residual       1163           1487.37736            1.27891

F =     371.65109       Signif F =  .0000

        * * * *  M U L T I P L E   R E G R E S S I O N  * * * *

Equation Number 1    Dependent Variable..   YQ

------------------ Variables in the Equation ------------------

Variable              B         SE B       Beta          T  Sig T

ZQ              .169773      .070895    .053699      2.395  .0168
X2Q            2.737096      .083530    .689449     32.768  .0000
X1Q             .135327      .041336    .073687      3.274  .0011
(Constant)     2.902456      .142225                20.407  .0000

Equation Number 1    Dependent Variable..   YQ
From Equation   1:   1 new variables have been created.

  Name        Contents
  ----        --------
  BRESID       Residual
```

Ergebnisse der Between-Regression

Nach der Neuanordnung und anschließender Aggregierung wird zunächst die Between-Regression über die Mittelwerte der Fälle gerechnet, für die alle drei Beobachtungen vorliegen. Die Ausgabe von SPSS/PC ist als Abbildung wiedergegeben. Die eigentlichen Schätzergebnisse sind für unsere Zwecke uninteressant. Von Bedeutung ist nur, daß die Residuen in der Variable "BRESID" gespeichert werden. Diese Residuen werden quadriert

und deren Summe (Variable: SSEB) zusammen mit der zugrundeliegenden Fallzahl (Variable: NBETW) in der Datei "tempb.sys" festgehalten.

Anschließend werden die Mittelwerte mit den Ursprungsdaten zusammengeführt. Nach dem Abziehen der Mittelwerte von den Ausgangsdaten wird die Within-Regression über die mittelwertsbereinigten Daten berechnet. Auch für diese Regression ist die Ausgabe von SPSS/PC wiedergegeben.

```
Page  35   EGLS-Schätzung des 'Error-Component-Models' (EC-Modell)      5/7/92

within-Regression (ohne Konstante!)

        * * * *  MULTIPLE REGRESSION THROUGH THE ORIGIN  * * * *

Equation Number 1    Dependent Variable..   YW

Block Number  1.  Method:  Enter
Variable(s) Entered on Step Number
   1..     X2W
   2..     X1W

Analysis of Variance
                    DF       Sum of Squares      Mean Square
Regression           2           354.93665        177.46832
Residual          5146          4261.89669           .82820

F =     214.28300       Signif F =  .0000

------------------ Variables in the Equation ------------------

Variable             B         SE B       Beta          T  Sig T

X2W              .986420     .047717    .277419    20.672  .0000
X1W              .005376     .028149    .002563      .191  .8485


From Equation   1:   1 new variables have been created.

  Name       Contents
  ----       --------
  WRESID      Residual
```

Ergebnisse der Within-Regression

Erklärende Variablen der Within-Regression sind allein die zeitvariablen Prädiktoren, im Beispiel also das politische Interesse und die Beurteilung der Notwendigkeit der Zählung. Da keine Regressionskonstante geschätzt wird, ist der von SPSS ausgegebenen Determinationskoeffizient nicht aussagekräftig (und in der Abbildung deswegen nicht wiedergegeben). Das Programm weist darauf auch durch einen kurzen Text hin.

Die geschätzten Regressionskoeffizienten sind dagegen brauchbare Schätzungen des Effektes der zeitvariablen Prädiktoren. Die ausgegebenen Standardfehler, T- und F-Test sind allerdings vor einer Interpretation zu korrigieren. Dies liegt daran, daß SPSS bei der Schätzung der Residualvarianz der Within-Regression die Summe der quadrierten Residu-

en durch die Anzahl der Beobachtungen minus der Anzahl der geschätzten Koeffizienten teilt. Korrekt wäre hier dagegen die Berechnung nach Gleichung (8). Korrigiert man die Werte der Standardschätzfehler um einen Korrekturfaktor q:

(15)
$$q = \sqrt{\frac{\sum_i T_i - p}{\sum_i T_i - n - p}}$$

erhält man korrekte Ergebnisse. Im Beispiel würde sich q nach

$$q = \sqrt{\frac{5148-2}{5148-2514-2}} = \sqrt{1.955} = 1.398$$

berechnen, da insgesamt 5148 Beobachtungen von 2514 verschiedenen Fällen vorliegen und zwei Regressionskoeffizienten in der Within-Regression geschätzt werden. Durch Multiplikation des Korrekturfaktors mit dem von SPSS ausgegebenen Wert erhält man die korrekten Standardfehler, im Beispiel den Wert 0.067 für den Standardfehler des Effekts der Notwendigkeit der Volkszählung (Variable: X2W) und 0.039 für den Standardfehler des Effekts des politischen Interesses (Variable: X1W). Die korrekten T-Werte erhält man, wenn die ausgegebenen Werte durch die Korrekturgröße geteilt werden. Im Beispiel sind die korrekten Werte 14.8 für die Notwendigkeit der Zählung und 0.14 für das politische Interesse. Auch der F-Test des Gesamteinflusses der Prädiktoren läßt sich korrigieren. Der von SPSS ausgegebene Wert (hier: 214.28) muß durch das Quadrat von q geteilt werden. Für die Beispieldaten erhält man einen Wert von 109.6. Die Freiheitsgrade sind 2 und 2632 (=5148-2514-2).

Für die eigentliche EGLS-Schätzung ist allein die Summe der quadrierten Residuen der Within-Regression von Bedeutung. Zur Berechnung werden die Residuen unter dem Variablennamen "WRESID" abgespeichert, quadriert und aufsummiert (Variable: SSEW). Das Ergebnis wird zusammen mit der Gesamtbeobachtungszahl (hier 5148) in der Datei "tempw.sys" festgehalten. Für die Berechnung der Varianzen der Fehlerkomponenten wird diese Datei mit den Dateien "tempb.sys", die die Ergebnisse der Between-Regression enthält, und "temp2.sys", die die Fallzahl enthält, zusammengefügt. Anschließend werden nach Gleichung (9) und (11) die Fehlervarianzen berechnet und ausgegeben.

```
Page  47   EGLS-Schätzung des 'Error-Component-Models' (EC-Modell)      5/7/92
***** Varianzen der Fehlerkomponenten:
  NFALL    NGES   NBETW      SIGMA2E      SIGMA2D   NEGVAR
   2514    5148    1167      1.61926       .73916        .
```

Für die zeitkonstante Komponente σ_δ^2 ergibt sich bei den Daten ein Schätzwert von 0.739, für die zeitveränderliche Komponente σ_ε^2 ein Wert von 1.619. Ausgegeben werden zusätzlich die Gesamtbeobachtungszahl (NGES $\hat{=} \Sigma T_i$), die Fallzahl (NFALL $\hat{=} n$), die Anzahl der Fälle der Between-Regression (NBETW) und eine Indikatorvariable NEGVAR, die den Wert 1 annehmen würde, wenn für σ_δ^2 ein negativer Wert geschätzt worden wäre.[7]

Die eigentliche Begründung für das EC-Modell liegt in der Gefahr der Autokorrelationen der Residuen aufgrund der Meßwiederholungen über die gleichen Untersuchungseinheiten. Aus den beiden Varianzkomponenten läßt sich diese Autokorrelation schätzen:

(16)
$$\hat{\rho} = \frac{\hat{\sigma}_\delta^2}{\hat{\sigma}_\delta^2 + \hat{\sigma}_\varepsilon^2}$$

Für das Beispiel ergibt sich ein Wert von 0.313. Da dieser Wert deutlich von Null abweicht, kann die Autokorrelation nicht einfach ignoriert werden. Die Schätzung der Regressionskoeffizienten nach Gleichung (3) über eine einfache OLS-Regression würde tatsächlich zu verzerrten Standardfehlern und Testergebnissen führen.

Nachdem die Varianzen der Fehlerkomponenten bestimmt sind, wird die eigentliche EGLS-Lösung berechnet. Dazu wird zunächst die Korrekturvariable $\hat{\vartheta}$ berechnet, dann die Ausgangsdaten (einschließlich der Konstante des Regressionsmodells) transformiert und schließlich die Kleinstquadratlösung der transformierten Daten berechnet und ausgegeben. Die vom Programm berechneten Regressionskoeffizienten, Standardfehler, T- und F-Werte sind interpretierbar. Der Ausschnitt aus der SPSS/PC-Ausgabe zeigt, daß die Beurteilung der Notwendigkeit einen deutlichen Einfluß auf die Haltung zur Volkszählung hat.

---

[7] Falls die Differenz aus Gleichung (11) negativ wird, wird der Schätzwert auf Null umkodiert, um das Programm fehlerfrei weiterrechnen zu lassen.

```
Page  51   EGLS-Schätzung des 'Error-Component-Models' (EC-Modell)      5/7/92

EGLS-Regress. über OLS der transform. Daten

        * * * *  MULTIPLE REGRESSION THROUGH THE ORIGIN  * * * *

Equation Number 1    Dependent Variable..   YT

Variable(s) Entered on Step Number
   1..     ZT
   2..     X2T
   3..     X1T
   4..     CONSTT

Analysis of Variance
                    DF       Sum of Squares       Mean Square
Regression           4         52425.39716       13106.34929
Residual          5144          9353.02203           1.81824

F =    7208.26494       Signif F =  .0000

------------------ Variables in the Equation ------------------

Variable             B         SE B        Beta           T  Sig T

ZT               .070629     .055949     .009795       1.262  .2069
X2T             1.990164     .046804     .295345      42.521  .0000
X1T              .088032     .024252     .056697       3.630  .0003
CONSTT          3.212612     .090192     .647564      35.620  .0000
```

Ergebnisse der EGLS-Schätzung

Befragte, die die Zählung für notwendig halten, weisen im Durchschnitt eine um 1.99 Einheiten positivere Haltung zur Volkszählung auf als Personen, die die Zählung nicht für notwendig halten. Der Effekt des politischen Interesses ist dagegen recht gering. Da ein steigender Wert der Variable für geringeres politisches Interesse steht, bedeutet das Ergebnis, daß sinkendes politisches Interesse mit einer geringfügig positiveren Bewertung der Volkszählung einhergeht. Das Geschlecht des Befragten hat schließlich keinen signifikanten Effekt auf die Haltung zur Zählung.

Von Interesse ist oft der Gesamteffekt der erklärenden Variablen. Berechnet man den Determinationskoeffizienten nach der Standardformel:

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y}_i)^2} = 1 - \frac{\sum_i e_i^2}{\sum_i (y_i - \bar{y})^2}$$

stellt sich die Frage, welche Residuen e; in die Formel eingehen sollen. Es liegt nahe, die Schätzungen von u_{it} aus Gleichung (4) zu verwenden. Diese liegen jedoch zunächst nicht vor, da die EGLS-Lösung auf der Basis der nach Gleichung (13) transformierten Daten berechnet wurde. Sie lassen sich jedoch berechnen, indem man von den Ausgangswerten der abhängigen Variable die Vorhersagewerte abzieht, die sich aufgrund der mit EGLS geschätzten Regressionskoeffizienten ergeben:

(17) $$\hat{u}_{it} = y_{it} - \hat{\alpha} - \sum_{j=1}^{p} x_{jit} \cdot \hat{\beta}_j - \sum_{l=1}^{k} z_{li} \cdot \hat{\gamma}_l$$

Etwas einfacher ist die Berechnung, wenn man die Summe der geschätzten Varianzen der beiden Fehlerkomponenten geteilt durch die geschätzte Varianz der abhängigen Variable von 1 abzieht:

(18) $$R^2_{adj} = 1 - \frac{\hat{\sigma}^2_\delta + \hat{\sigma}^2_\varepsilon}{\hat{\sigma}^2_y}$$

Da bei der Berechnung nach Gleichung 18 die Freiheitsgrade angepaßt sind, handelt es sich um einen adjustierten Determinationskoeffizient. Für das Beispiel berechnet sich der Wert nach:

$$R^2_{adj} = 1 - \frac{1.62 + 0.74}{3.94} = 0.40$$

Die drei erklärenden Variablen decken demnach ungefähr 40% der Varianz der abhängigen Variable auf.

5. Diskussion

Das Beispiel zeigt, daß die EGLS-Schätzung der Koeffizienten des EC-Modells auch mit SPSS/PC möglich ist. Für eine halbwegs automatische Berechnung ist allerdings eine Vielzahl von Anweisungen möglich. Einmal geschrieben, können sie aber jederzeit als Gerüst für weitere Anwendungen verwendet werden. Sollen etwa zusätzliche Variablen verwendet werden, ist es nur nötig, bei dem im Anhang dokumentierten Programm in den SAVE-, AGGREGATE-, JOIN ADD-, JOIN MATCH- und REGRESSION-Anweisungen zusätzliche Variablen anzuhängen und die Konstanten "p" bzw. "k" entsprechend hochzusetzen. Bei Bedarf kann auch die Konstante "tmax" angepaßt werden, wenn bei der Between-Regression Fälle mit einer anderen Beobachtungshäufigkeit berücksichtigt werden sollen.

Möglich ist ferner die Berücksichtigung von Dummy-Variablen für den jeweiligen Meßzeitpunkt. Die Spezifikation solcher Dummy's bedeutet inhaltlich, daß davon ausgegangen wird, daß das Durchschnittsniveau der abhängigen Variablen unabhängig von möglichen Änderungen bei den zeitveränderlichen Prädiktoren zu jedem Meßzeitpunkt unterschiedlich sein kann. So haben andere Analysen der Daten der Kölner Begleituntersuchung gezeigt,

daß sich zwischen der ersten und der zweiten Panelwelle ein deutlicher Wandel zugunsten einer positiveren Einstellung zur Volkszählung ereignete (vgl. *Scheuch u.a.,* 1989). Diese Veränderung läßt sich dadurch spezifizieren, daß eine Dummy-Variable als weiterer Prädiktor ins Modell aufgenommen wird, die in der ersten Panelwelle den Wert Null und in den Folgewellen den Wert Eins aufweist. Wenn die Einstellungsänderung nicht bereits durch eine Änderung der Beurteilung der Notwendigkeit der Zählung erklärt werden kann, wird diese Dummy-Variable einen signifikanten positiven Effekt aufweisen.

Bei der Spezifikation solcher Dummy-Variablen ist eine gewisse Vorsicht angebracht, da die Dummy-Variablen weder reine zeitvariable noch zeitkonstante Variablen sind. Bei der Berechnung der Mittelwerte über die Zeit ergibt sich nämlich bei allen Untersuchungseinheiten mit gleichem Beobachtungsmuster der gleiche Wert. Bei der Between-Regression lassen sich daher diese Effekte nicht schätzen. Auch bei der Within-Regression kann es Schwierigkeiten, geben, wenn für alle Untersuchungseinheiten die gleiche Anzahl von Meßzeitpunkten vorliegt und mehr als eine Dummy-Variable spezifiziert wird. Am einfachsten ist es, wenn die Dummy-Variablen bei der Schätzung der Varianzkomponenten nicht berücksichtigt werden und erst im letzten Schritt in die Regressionsgleichung aufgenommen werden.[8]

So einfach Analysen mit dem EC-Modell sind, so sei doch auch auf Probleme hingewiesen. Sichtbar werden diese, wenn man die Koeffizientenschätzung der Within-Regression mit der EGLS-Lösung vergleicht. Theoretisch handelt es sich jeweils um konsistente Schätzungen derselben Koeffizienten. Trotzdem weisen die Schätzwerte erhebliche Unterschiede auf.[9] Woran mag dies liegen? Der wichtigste Unterschied zwischen den beiden Schätzverfahren besteht darin, daß die Effekte der Within-Regression allein durch die Veränderungen bei den Werten der zeitveränderlichen Variablen zu verschiedenen Meßzeitpunkten induziert werden. Insbesondere bei dem politischen Interesse ist es aber denkbar, daß dieses genaugenommen zeitkonstant ist, Veränderungen also alleine durch mangelnde Reliabilität hervorgerufen werden. Wenn dem so wäre, würden bei der Within-Regression nur Meßfehler in die Schätzung des Effekts dieser Variable eingehen. Tatsächlich wird bei der Within-Regression wie bei allen linearen Regressionsmodellen für beobachtbare Variablen unterstellt, daß die erklärenden Variablen meßfehlerfrei erfaßt werden. Ist dies nicht der Fall, können die Ergebnisse sehr stark verzerrt sein.

---

[8] Vor der Schätzung der Regressionskoeffizienten müssen allerdings auch die Werte der Dummy-Variablen nach Gleichung (13) transformiert werden.

[9] Die Unterschiede zwischen der Within- und der EGLS-Schätzung kann in statistischen Tests ausgenutzt werden, die in gewisser Hinsicht die Angemessenheit des spezifizierten EC-Modells testen (vgl. *Hübler,* 1990: 72f.). Unglücklicherweise ist bei einem signifikanten Testergebnis i.a. nicht sicher, welche Annahmen denn nun tatsächlich verletzt sind.

Falls also mit einem erheblichen Anteil von Meßfehlern zu rechnen ist, sollte statt des EC-Modells für beobachtete Daten auf Modelle zurückgegriffen werden, die solche Meßfehler berücksichtigen können. Entsprechende Modelle werden in der Arbeit von ***Arminger und Müller*** (1990) diskutiert. Die größere Komplexität jener Modelle bringt es jedoch mit sich, daß sie nicht mehr so einfach mit SPSS geschätzt werden können wie das hier behandelte EC-Modell. Als Strategie bietet sich daher möglicherweise an, zunächst mit einfachen Modellen wie dem EC-Modell zu beginnen und erst in einem späteren Analysestadium auf komplexere Modelle umzusteigen.

6. Literatur

Arminger, G. und F. Müller (1990), Lineare Modelle zur Analyse von Paneldaten. Opladen: Westdeutscher Verlag.

Hübler, O. (1990), Lineare Panelmodelle mit alternativer Störgrößenstruktur. **In:** ***G. Nakhaeizadeh und K.-H. Vollmer*** (Hrsg.), Neuere Entwicklungen in der Angewandten Ökonometrie. Beiträge zum 1. Karlsruher Ökonometrie-Workshop. Heidelberg: Physica-Verlag.

Rendtel, U. (1992), Handouts zur Vorlesung "Panelanalyse" beim Frühjahrsseminar 1992 des Zentralarchiv für empirische Sozialforschung, Köln.

Scheuch, E.K., Graf, L. und S. Kühnel (1989), Volkszählung, Volkszählungsprotest und Bürgerverhalten. Ergebnisse der Begleituntersuchung zur Volkszählung 1987. Stuttgart: Metzler-Poeschel.

ANHANG: Dokumentation der SPSS/PC-Anweisungen

```
set more off.
*** EGLS-Schätzung des EC-Modells mit SPSS/PC (V4.01) ***.
TITLE "EGLS-Schätzung des 'Error-Component-Models' (EC-Modell)".

** Vorbereitung: Re-Arrangieren der Paneldaten
   Die zu analysierende Datei "ecbeisp.sys" enthält die Variablen:
   IDNR      Fallnummer
   Y1        abhäng. Variable, 1. Panelwelle
   Y2        abhäng. Variable, 2. Panelwelle
   Y3        abhäng. Variable, 3. Panelwelle
   X11       zeitabh. Prädiktor X1, 1. Welle
   X12       zeitabh. Prädiktor X1, 2. Welle
   X13       zeitabh. Prädiktor X1, 3. Welle
   X21       zeitabh. Prädiktor X2, 1. Welle
   X22       zeitabh. Prädiktor X2, 2. Welle
   X23       zeitabh. Prädiktor X2, 3. Welle
   Z         zeitkonstanter Prädiktor.
SUBTITLE "Rearrangieren der Daten".
GET FILE="ecbeisp.sys".
* 1. Welle.
COUNT nmissing=y1 x11 x21 z (MISSING).
COMPUTE zeit=1.
PROCESS IF (nmissing EQ 0).
SAVE OUT="temp1.sys"
    /KEEP=idnr zeit y1 x11 x21 z /RENAME (y1 x11 x21=y x1 x2).
* 2. Welle.
GET FILE="ecbeisp.sys".
```

```
COUNT nmissing=y2 x12 x22 z (MISSING).
COMPUTE zeit=2.
PROCESS IF (nmissing EQ 0).
SAVE OUT="temp2.sys"
    /KEEP=idnr zeit y2 x12 x22 z /RENAME (y2 x12 x22=y x1 x2).
* 3. Welle.
GET FILE="ecbeisp.sys".
COUNT nmissing=y3 x13 x23 z (MISSING).
COMPUTE zeit=3.
PROCESS IF (nmissing EQ 0).
SAVE OUT="temp3.sys"
    /KEEP=idnr zeit y3 x13 x23 z /RENAME (y3 x13 x23=y x1 x2).
* Wiedereinlesen und sortieren der Fälle.
JOIN ADD /FILE="temp1.sys" /FILE="temp2.sys" /FILE="temp3.sys".
SORT CASES BY idnr zeit.
* Festlegung von Konstanten:
   p:    Anzahl zeitveränderlicher Prädiktoren (hier: 2)
   k:    Anzahl zeitkonstanter Prädiktoren (hier: 1)
   tmax: maximale Anzahl von Beobachtungen pro Fall über die Zeit (hier 3).
COMPUTE p=2.
COMPUTE k=1.
COMPUTE tmax=3.
* Zwischenspeichern der Ausgangsdaten.
SAVE OUT="temp1.sys" /COMP.

***** Schritt 1: Schätzen der Varianzen der Fehlerkomponenten *****.
** Berechnung der Mittelwerte über die Zeit.
SUBTITLE "Berechnung der Mittelwerte über Zeit durch Aggregation".
* die Mittelwerte über die Zeit sind durch ein 'q' nach dem Variablennamen
  gekennzeichnet, also  'yq' für 'Mittelwert von Variable y'.
* Auch die Konstanten p,k und tmax müssen in die Datei mit den
  aggregierten Fällen übertragen werden!.
AGGREGATE OUT=*
    /PRES /BREAK=IDNR
    /yq=MEAN(y)
    /x1q=MEAN(x1)
    /x2q=MEAN(x2)
    /zq=MEAN(z)
    /ti=NU(y)
    /p=MEAN(p)
    /k=MEAN(k)
    /tmax=MEAN(tmax).
* Berechnung der Hilfsgröße 'const' und Speicherung der Mittelwerte
  in Datei 'tempm.sys'.
COMPUTE const=1.
SAVE OUT="tempm.sys" /COMP.

**  Zwischenspeichern der Fallzahl in 'temp2.sys".
AGGREGATE OUT="temp2.sys" /BREAK=const /NFALL=NU(idnr).

**  Regression über die Mittelwerte: Between-Regression über vollständige
    Fälle; Sichern der Residuen in Variable 'bresid'.
SUBTITLE "Berechnung 'between'-Regression' über vollständige Fälle".
SELECT IF (ti=tmax).
REGRESSION VAR yq x1q x2q zq
    /DEP yq
    /ENT
    /SAVE=RESID(bresid).

** Berechnung und Speicherung der quadrierten Between-Residuen und der
   Fallzahl der Regressions in die Variablen 'sseb' und 'nbetw' in Datei
   'tempb.sys'.
SUBTITLE "Berechnung u. Speicherung SSE-'beetween' und Fallzahl".
COMPUTE bres2=bresid*bresid.
AGGREGATE OUT="tempb.sys" /PRES /BREAK=const
 /nbetw=NU(idnr)
 /sseb=SUM(bres2)
```

```
 /p=MEAN(p)
 /k=MEAN(k)
 /tmax=MEAN(tmax).

** Schätzung within-Regression.
*  Die mittelwertsbereinigten Daten sind durch ein 'w' hinter den
   Variablennamen gekennzeichnet, also 'yw' für die mittelwertsbereinigte
   Variable 'y'.
SUBTITLE "Mittelwertsbereinigung für 'within'-Regression".
JOIN MATCH  TABLE="tempm.sys"
          /FILE="temp1.sys"
          /BY idnr.
COMPUTE yw=y-yq.
COMPUTE x1w=x1-x1q.
COMPUTE x2w=x2-x2q.
*  within-Regression.
*  Die Residuen werden in Variable 'wresid' festgehalten.
SUBTITLE "within-Regression (ohne Konstante!)".
REGRESSION VAR yw x1w x2w
         /ORIGIN /DEP=yw /ENT
         /SAVE=RESID(wresid).

** Speicherung der quadrierten Residuensumme ('ssew') und der Anzahl
   Beobachtungen (über Fälle und Zeit: 'nges') in Datei 'tempw.sys'.
SUBTITLE "Berechnung u. Speicherung SSE-'within' und Beobachtungszahl".
COMPUTE wres2=wresid*wresid.
SAVE OUT="temp1.sys" /COMP.
AGGREGATE OUT="tempw.sys" /PRES /BREAK=const
  /nges=NU(idnr)
  /ssew=SUM(wres2).

** Berechnung der Varianzen 'sigma2e' und 'sigma2d' der Fehlerkomponenten.
SUBTITLE "Berechnung der geschätzten Fehlerkomponenten".
JOIN MATCH FILE="tempb.sys" /FILE="tempw.sys" /FILE="temp2.sys".
COMPUTE sigma2e=ssew/(nges-nfall-p).
COMPUTE sigma2d=sseb/(nbetw-p-k-1) - sigma2e/tmax.
* Nullsetzen von 'sigma2d' bei negativer Schätzung.
IF (sigma2d LT 0) negvar=1.
IF (sigma2d LT 0) sigma2d=0.
FORMATS sigma2e sigma2d (F12.5).
SUBTITLE "***** Varianzen der Fehlerkomponenten:".
LIST VARIABLES= nfall nges nbetw sigma2e sigma2d negvar.

***** Schritt 2: Berechnung der EGLS-Lösung über OLS transformierter Daten.
** Berechnung der Transformationsformel 'theta'.
SUBTITLE "Berechnung von 'theta'".
JOIN MATCH TABLE=*   /FILE="tempm.sys" /BY const.
COMPUTE theta=1-SQRT(sigma2e/(sigma2e+sigma2d*ti)).

** Transformation der Ursprungsdaten und EGLS-Regression.
*  Die transformierten Daten sind durch ein 't' nach dem Variablennamen
   gekennzeichnet, so 'constt' für die transformierte Konstante.
SUBTITLE "EGLS-Regress. über OLS der transform. Daten".
JOIN MATCH TABLE=* /FILE="temp1.sys" /BY idnr.
COMPUTE constt=1-theta.
COMPUTE yt=y-theta*yq.
COMPUTE x1t=x1-theta*x1q.
COMPUTE x2t=x2-theta*x2q.
COMPUTE zt=z-theta*z.
REGRESSION VAR=yt constt x1t x2t zt
   /ORIGIN /DEP=yt /ENT.
*****************************************************************.
```